第 38 卷第 3 期
录用定稿
计算机应用研究
Application Research of Computers
Vol. 38 No. 3
Accepted Paper

基于项目模糊相似度的协同过滤推荐算法

王　森，陈　莉†，张　洁

(西北大学 信息科学与技术学院，西安 710127)

摘　要：针对传统协同过滤算法中评分和标签存在的模糊性问题进行了研究，利用梯形模糊数描述评分与满意度的映射关系，在考虑评分稀疏性的基础上构建了一种新的梯形模糊评分模型以判断基于模糊评分的相似度，分析标签与项目的隶属度，构建模糊项目标签矩阵以衡量基于标签隶属度的相似度，最终采用改进的评分预测策略进行评分估计。在 MovieLens 数据集上的实验结果显示，所提算法在抑制项目冷启动、缓解模糊性和稀疏性问题的同时，提高了预测精度，表明了该算法的有效性。

关键词：协同过滤；模糊相似度；梯形模糊评分模型；模糊项目标签矩阵

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2020.04.0056

Collaborative filtering recommendation algorithm based on item fuzzy similarity

Wang Sen, Chen Li†, Zhang Jie

(*School of Information Science & Technology*, *Northwest University*, *Xi'an* 710127, *China*)

Abstract: In view of the problem of fuzziness of rating and tag in traditional collaborative filtering algorithms, a trapezoidal fuzzy number was used to describe the mapping relationship between rating and satisfaction. The algorithm considered the impact of sparseness of the rating, constructed a new trapezoidal fuzzy rating model to determine the similarity based on fuzzy rating, analyzed the degree of membership between the tag and the item, and constructed a fuzzy item-tag matrix to measure the similarity based on the degree of tag membership. Finally, the improved scoring prediction strategy was used to estimate the score. The experimental results on the MovieLens dataset show that the proposed algorithm improves the prediction accuracy while suppressing the cold start of the project, alleviating the problems of fuzziness and sparseness, which indicate the effectiveness of the proposed algorithm.

Key words: collaborative filtering; fuzzy similarity; trapezoidal fuzzy rating model; fuzzy item-tag matrix

0　引言

伴随着信息技术的蓬勃发展，机构和个人用户产生的数据量急剧增加，导致 WEB 用户难以高效获取有价值的信息[1]。推荐系统因其可以主动预测用户需求，为用户推荐数据的特点，已成为缓解信息过载问题最常用的方法之一。

协同过滤(collaborative filtering, CF)是目前应用最广泛的推荐技术，其借助“物以类聚、人以群分”的思想，认为用户对项目的偏好可以根据邻域的其余用户对项目的评价进行推测，或者根据用户对目标项目邻域的评价进行推测[2]。但是，由于互联网数据的急剧增加和用户习惯的缺陷，协同过滤算法依然面临着稀疏性、冷启动等问题。

学者们针对以上问题展开了研究，例如，Wei 等人[3]将时间感知协同过滤模型 timeSVD ++与深度学习架构 SDAE 相结合，利用 SDAE 提取项目内容特征，使用 timeSVD ++模型预测评分以解决项目冷启动问题。Hu 等人[4]提出了一种相似度增强机制，通过中间用户和项目发掘潜在的相似性关系，并从相似邻居中提取更多数据以减少稀疏性问题。但是此类研究大多未考虑协同过滤的模糊性问题[5~7]，影响了推荐质量。

模糊性问题是指协同过滤算法的输入通常具有模糊性。例如在电影评分系统中，用户需要从给定评分集合中选择某评分以表达自己对项目的满意度，但是有限的评分不能充分表达用户的偏好差异，往往只能选择一个接近自己喜好程度的评分，这就意味着相同评分并不代表用户的偏好完全一致。另外，由于时间、心情和环境等因素的影响，相同的满意度也可能表现为不同的评分。这种评分-满意度关系的不确定性被称为评分模糊性。此外，推荐系统中还存在项目标签隶属度不确定等模糊性问题。

针对该问题，学者们提出了模糊协同过滤算法，使用模糊理论更恰当地表达协同过滤算法输入数据蕴涵的信息，提升预测精度。Tsai 等人[8]采用模糊集度量两个业务之间的相似度，以预测两家企业吸引同一用户的可能性。该算法的准确率高于对比算法 33%，体现了算法的优势。但是其所构建的评论网络仅仅是基于同一个用户对两家企业发表评论的二进制变量，未考虑评论的方差，有进一步改进预测模型的必要。Vashisth P 等人[9]使用区间 2 型模糊集创建用户模型，以捕捉不同用户行为的模糊性，基于此提出了模糊特征混合方法(fuzzy feature combination hybridzation method, FFCHM)，改善了数据稀疏性问题，但是该方法的时间复杂度过高，可扩展性低。Wasid M 等人[10]针对基于内存的协同过滤的可扩展性问题，提出了一种基于用户模糊特征的推荐系统，他们认为大多数用户特征在本质上是模糊的，使用模糊集可以更精确地描述用户特征。该算法相比于传统协同过滤算法在 MAE、覆盖率、准确性和效率等指标上都有提升。Kant 等人[11]使用局部模糊距离和全局模糊距离衡量用户和项目相似度以提升预测精度。实验结果表明该算法的覆盖率较高、MAE 值较低，但是在稀疏数据集中的表现不佳。Zhang X 等人[12]使用三角模糊数描述用户对项目的综合评价，根据三角形面积和中点衡量三角模糊数相似度以确定用户相似度，提升了相似度计算的准确率。然而三角模糊数中隶属度的最大

收稿日期：2020-04-08；**修回日期：**2020-05-23

作者简介：王森(1995-)，男，陕西咸阳人，硕士研究生，主要研究方向为个性化推荐；陈莉(1963-)，女(通信作者)，陕西西安人，教授，博导，博士，主要研究方向为智能信息处理、数据挖掘、网络安全(chenli@nwu.edu.cn)；张洁(1995-)，女，陕西西安人，硕士研究生，主要研究方向为计算机视觉．

值只对应一个点，灵活性逊于梯形模糊数，导致可扩展性较低。吴毅涛等人[13]借鉴年龄模糊模型将满意度映射到原始评分上，引入梯形模糊相似度计算策略衡量用户相似度以提升推荐效果。他们使用数学方式证明了模糊相似度是余弦相似度在模糊域上的扩展，实验结果表明该算法的预测精度优于基于三角模糊数的协同过滤算法。但是该算法的模型是固定的，无法随着数据集和用户的改变而自动调整。2017 年，吴毅涛等人[14]在文献[13]的基础上，根据评分的分布情况自动生成个性化梯形模糊评分模型，基于模糊相似度和模糊评分实施评分预测以改进推荐质量，实验结果表明该算法的预测误差较低。但是该算法的模型未考虑评分数据的稀疏性问题，致使部分低频率评分对应的梯形模糊数误差较大。

综上所述，模糊协同过滤算法可降低输入数据的模糊性，提高相似度计算的准确率，提升推荐质量。但是，目前大多数模糊协同过滤算法依然存在以下几个方面的问题。

a) 只模糊化协同过滤的部分过程，缺少对数据预处理、相似性计算和评分预测等全过程实施模糊化的算法。

b) 模糊数的相似度计算只考虑常规距离和重心距离，误差仍然较大。

c) 忽略数据稀疏性对模糊化准确率的影响。

针对以上不足，本文改进稀疏数据导致的评分统计噪声的问题，将评分转换成梯形模糊数，使用新的梯形模糊相似度计算策略判断基于模糊评分的项目相似性(item similarity based on fuzzy rating, FRIS)，利用模糊隶属度将标签与项目的关系由{0,1}扩展为[0,1]，并以此判断基于标签隶属度的项目相似度(item similarity based on tags membership, TMIS)，然后融合以上两种相似度形成项目相似度，使用一种新的模糊评分预测策略进行评分估计，最终用于推荐。在 MovieLens100K 和 1M 数据集上进行实验，结果表明本文算法可在一定程度上改善模糊性和稀疏性带来的不利影响，抑制项目冷启动问题。

1 相关工作

1.1 模糊集合与模糊数

经典集合论中元素 i 与集合 U 的关系只能是属于或者不属于，且满足式(1)，其中 ⊕ 表示异或关系。

$$[(i \in U) \oplus (i \notin U)] = 1 \tag{1}$$

然而现实世界中的概念大多不是非此即彼的。例如交通工具的时速，有人认为 60km/h 的速度快，另一部分人认为慢。

针对以上问题，L.A.Zadeh 教授[15]提出了模糊理论，使用数学工具描述客观世界的模糊现象，利用隶属度函数将二值逻辑改进为连续值逻辑。

模糊集合是模糊理论的数据表现形式，若给定论域 U，集合 A 有式(2)所示的映射关系，则称 A 是模糊集合。

$$\mu_A: U \to [0,1],\ u \mapsto \mu_A(u) \tag{2}$$

μ_A 是 A 的隶属度函数，$\mu_A(u)$ 的范围是[0,1]。若将模糊集合的隶属度 $\mu_A(u) \in [0,1]$ 变为 $\mu_A(u) \in \{0,1\}$，则模糊集合退化为经典集合。

模糊数是满足特定要求的模糊集合，可以更精确地表现协同过滤的映射关系，同时也方便数学处理，模糊数的相关概念如定义 1 和 2 所示。

定义 1 假设 $A \in F(R)$，对于 $\forall \lambda \in (0,1]$，称 $A_\lambda = \{u \in U \mid \mu_A(u) \ge \lambda\}$ 为 A 的 λ 截集，λ 为置信度，其中 $A \in F(R)$ 表明 A 是实数集 R 上的模糊集。

定义 2 假设 $A \in F(R)$，$\exists x \in R$ 使得 $\mu_A(x) = 1$，且 $\forall \lambda \in (0,1]$，$A_\lambda$ 是闭区间，则称 A 是模糊数。

1.2 梯形隶属度函数

隶属度函数决定了对象与模糊集合之间的隶属度，可用来描述模糊集合，目前常用的隶属度函数有三角形、梯形、高斯型和钟型等函数。

三角形与高斯型隶属度函数的趋势类似，隶属度先升高再降低，区别在于高斯型的曲线可以更精确的描述模糊概念。这两种函数适用于描述某一明确定义的附近范围，在明确定义处，隶属度最大。但由于以上两种函数只允许一点处的隶属度为最大值，故不符合用户评分的习惯。

梯形与钟型隶属度函数的趋势类似，隶属度先升高再保持最后降低，但是由于钟型隶属度函数构建模型的时间消耗过大，故可行性较低。梯形隶属度函数因为容易操作、计算简单且比较贴合大多数模糊概念的特点，应用场景最为广泛。并且，梯形隶属度函数可以通过调整参数退化成三角隶属度函数以扩大适用范围，可扩展性强。所以，本文拟使用梯形隶属度函数改进满意度与评分的映射关系，以提升推荐效果。梯形模糊隶属度函数如图 1 所示，定义如式(3)所示。

$$Trap(x;a,b,c,d)=\begin{cases}0, x \le a \\ \dfrac{x-a}{b-a}, a \le x \le b \\ 1, b \le x \le c \\ \dfrac{d-x}{d-c}, c \le x \le d \\ 0, d \le x\end{cases} \tag{3}$$

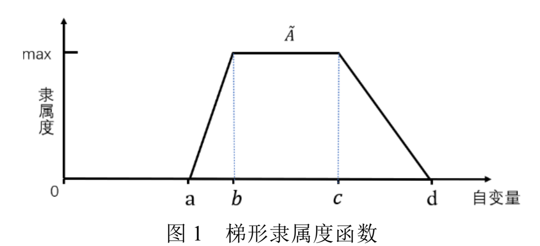


图 1 梯形隶属度函数

Fig. 1 Trapezoid membership function

1.3 梯形模糊数的运算

使用梯形隶属度函数的模糊数被称为梯形模糊数。梯形模糊数可以使用梯形四个顶点的横坐标和最大隶属度进行描述。具体定义如式(4)所示。

$$A_i = (a_{i,1}^A, a_{i,2}^A, a_{i,3}^A, a_{i,4}^A; W_i) \tag{4}$$

上式中 $a_{i,j}^A$ 表示梯形模糊数 A_i 的 j 个顶点的横坐标值，W_i 表示该梯形模糊评分的最大隶属度。

假设有两个梯形模糊数 A_i 和 A_k，则它们的加、减、乘、除运算如式(5)(6)(7)和(8)所示。

$$A_i + A_k = \begin{pmatrix} a_{i,1}^A + a_{k,1}^A, a_{i,2}^A + a_{k,2}^A, a_{i,3}^A + a_{k,3}^A, \\ a_{i,4}^A + a_{k,4}^A; W_i + W_k \end{pmatrix} \tag{5}$$

$$A_i - A_k = \begin{pmatrix} a_{i,1}^A - a_{k,1}^A, a_{i,2}^A - a_{k,2}^A, a_{i,3}^A - a_{k,3}^A, \\ a_{i,4}^A - a_{k,4}^A; W_i - W_k \end{pmatrix} \tag{6}$$

$$A_i * t = (a_{i,1}^A * t, a_{i,2}^A * t, a_{i,3}^A * t, a_{i,4}^A * t; W_i * t) \tag{7}$$

$$A_i / t = (a_{i,1}^A / t, a_{i,2}^A / t, a_{i,3}^A / t, a_{i,4}^A / t; W_i / t) \tag{8}$$

本文将结合以上公式，改进现有协同过滤系统的评分预测策略，利用邻域的梯形模糊评分预测目标评分。

1.4 信息量

香农在信息论中通过事件发生的不确定性度量事件包含的信息量，即事件发生的可能性与其包含的信息量成反比。信息量计算如式(9)所示。

$$H^i = -p^i * \log_2 p^i \tag{9}$$

其中，H^i 表示事件 i 的信息量，p^i 表示事件 i 发生的概率。

本文拟通过评分出现的概率计算评分信息量，以此作为权重调整不同项目对用户相似性的贡献度，以提升相似度计算的准确率。

2　基于项目模糊相似度的协同过滤推荐算法

传统的项目协同过滤算法忽略了输入数据的模糊性，致使推荐精度较低。已有的模糊协同过滤算法常采用经典的模糊相似度计算策略，导致相似度计算误差较大，且多是从用户角度出发，忽略了项目标签表达的信息，使得推荐效果欠佳。所以，基于项目的模糊协同过滤算法仍需进一步研究。

针对以上问题，本文对协同过滤算法的全过程进行模糊化，构建梯形模糊评分模型和模糊项目标签矩阵，提出一种新的模糊相似度计算策略以提升推荐质量。

2.1　一种新的梯形模糊评分模型

满意度是判断用户相似度的关键因素。因此，如何更好地描述用户对项目的满意度是协同过滤算法的关键问题之一。针对此问题，本节改进稀疏数据集的评分统计方法，以满意度作为论域，利用梯形隶属函数将满意度映射到评分中，以使用梯形模糊评分代替原始评分进行相似度计算。

目前的评分系统多为 5 评分或 10 评分系统，本节以 5 评分系统为例，介绍新的梯形模糊评分模型。

梯形模糊评分模型如图 2 所示，横坐标为满意度，纵坐标为隶属度，范围是[0,1]，纵坐标的值越大，表示满意度对应梯形模糊数的隶属度越大。因为任意满意度都可使用评分进行描述，故在该模型中，任意满意度对应的评分隶属度之和都为 1。

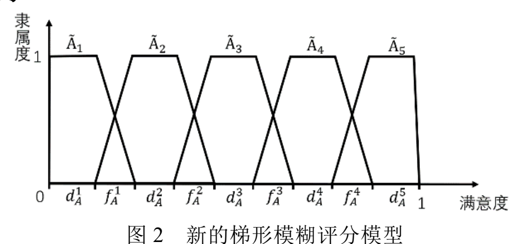


图 2　新的梯形模糊评分模型

Fig. 2　New trapezoidal fuzzy scoring model

图 2 中 A_i 是用户 A 的第 i 种梯形模糊评分，i 为原始评分，i∈{1,2,3,4,5}。A_i 对应的满意度置信区间由 f_A^{i-1}、d_A^i、f_A^i 三部分(i∈{2,3,4})或者其中两部分(i∈{1,5})组成。d_A^i 表示 A_i 对应满意度的确定域，在 d_A^i 范围内，满意度只映射到一个梯形模糊评分中。f_A^{i-1} 和 f_A^i 表示 A_i 对应满意度的模糊域，在该范围内，任意满意度都映射到两个相邻的梯形模糊评分中。

根据模糊统计法，用户 A 的第 i 种评分出现的概率决定 A_i 对应满意度置信区间的大小 r_A^i，评分 i 出现的概率越大说明用户越喜欢使用该评分评价项目，则评分 i 对应的满意度置信区间越大。r_A^i 的定义如式(10)所示。

$$r_A^i = \frac{n(r_A = i)}{n(r_A)} \tag{10}$$

其中，$n(r_A = i)$ 表示用户 A 的评分中 i 出现的次数，$n(r_A)$ 表示用户 A 的总评分数。

由于数据过度稀疏，故会出现用户未曾使用过某种评分的现象，导致该评分对应的满意度置信区间长度为 0。为了避免以上问题，本文使用 $new(r_A = i)$ 和 $new(r_A)$ 改进 $n(r_A = i)$ 和 $n(r_A)$，以改善数据稀疏性造成的统计误差，具体定义如式(11)和(12)所示。

$$new(r_A = i) = \begin{cases} n(r_A) * t, if\,(n(r_A = i) < n(r_A) * t) \\ n(r_A = i), if\,(n(r_A = i) \ge n(r_A) * t) \end{cases} \tag{11}$$

$$new(r_A) = \sum_{i=1}^{5} new(r_A = i) \tag{12}$$

式(11)中 $new(r_A = i)$ 表示改进后 A 用户评分 i 的出现次数，通过赋值可变参数 t 可以调整评分出现次数的最小值，进而调整评分 i 对应的最小置信区间长度。式(12)中 $new(r_A)$ 表示改进后用户 A 的总评分数。

通过以上改进，r_A^i 的定义如式(13)所示。r_A^i 表示用户 A 的评分 i 对应的满意度置信区间长度，故 $\sum_{i=1}^{5} r_A^i = 1$。

$$r_A^i = \frac{new(r_A = i)}{new(r_A)} \tag{13}$$

r_A^i 表示的满意度置信区间由模糊域和确定域组成，相关定义如式(14)和(15)所示。

$$f_A^i = 2 * p * \min(r_A^i, r_A^{i+1}) \tag{14}$$

$$d_A^i = \begin{cases} r_A^i - 0.5 * f_A^i - 0.5 * f_A^{i-1}, i \in \{2,3,4\} \\ r_A^i - 0.5 * f_A^i, i = 1 \\ r_A^i - 0.5 * f_A^{i-1}, i = 5 \end{cases} \tag{15}$$

f_A^i 表示模糊域，位于两个评分对应满意度的交界处，由 r_A^i 和 r_A^{i+1} 中的最小数与描述模糊程度的参数 p 确定。从图 2 可知 $f_A^i \ge 0$ 且 $f_A^i \le \min(r_A^i, r_A^{i+1})$，所以 $p \in [0, 0.5]$。d_A^i 表示确定域，长度由 r_A^i 与模糊域共同决定，该模型中每种评分对应一个确定域。综上所述，5 评分系统的梯形模糊评分模型由 4 个模糊域、5 个确定域构成，若将确定域的长度缩小为一点，则梯形模糊评分模型退化为三角模糊评分模型。

本模型中所有梯形模糊数的 1 和 4 顶点的纵坐标值都为 0，顶点 2、3 的纵坐标值都为 1，故本模型的梯形模糊数定义如式(16)所示。

$$A_i = (a_{i,1}^A, a_{i,2}^A, a_{i,3}^A, a_{i,4}^A, 1) \tag{16}$$

$$A_i = \begin{cases} (0, 0, r_A^1 - 0.5 * f_A^1, r_A^1 + 0.5 * f_A^1, 1), i = 1 \\ \begin{pmatrix} \sum_{j=1}^{i-1} r_A^j - 0.5 * f_A^{(i-1)}, \sum_{j=1}^{i-1} r_A^j + 0.5 * f_A^{(i-1)}, \\ \sum_{j=1}^{i} r_A^j - 0.5 * f_A^i, \sum_{j=1}^{i} r_A^j + 0.5 * f_A^i, 1 \end{pmatrix},\ i \in \{2,3,4\} \\ \left(\sum_{j=1}^{4} r_A^j - 0.5 * f_A^4, \sum_{j=1}^{4} r_A^j + 0.5 * f_A^4, 1, 1, 1 \right), i = 5 \end{cases} \tag{17}$$

式(16)中 $a_{(i,j)}^A$ 表示用户 A 的评分 i 第 j 个顶点的横坐标值，1 表示该梯形模糊评分的最大隶属度为 1。

结合模糊域与确定域的定义，A_i 的定义如式(17)所示。

2.2　梯形模糊相似度

协同过滤算法常用的余弦相似度等计算策略不适用于模糊评分的相似度计算，故本小节以梯形模糊相似度为基础，结合评分信息改进梯形模糊评分模型的相似度计算策略[16]。

设两个梯形模糊评分为 A_i 和 B_j，根据 Ahmad S 的定义，它们的相似度计算如式(18)(19)和(20)所示。

式(18)中，四个因子分别表示几何距离、重心距离、戴斯相似性系数(dice similarity coefficient，DSC)、豪斯多夫距离(hausdorff distance,HD)。其中，几何距离度量了两个梯形的横向距离，重心距离度量了梯形重心的横向与纵向距离，且由于重心纵坐标与梯形上下底的长度差距有关，故重心距离也反映了梯形模糊评分中模糊域和确定域的差别，DSC 通常根据两个集合的重复比例判断它们的相似性，在式(18)中，Ahmad S 将 DSC 引入梯形模糊评分，利用梯形顶点的横坐标判断梯形的相似性，HD 将顶点横坐标视为点集以判断梯形模糊评分的最大不匹配程度。

W_{A_i} 和 W_{B_j} 表示梯形模糊评分 A_i 和 B_j 的最大隶属度，本文中设定 $W_{A_i} = W_{B_j} = 1$。X_{A_i} 和 X_{B_j} 分别表示梯形模糊评分 A_i 和 B_j 重心的横坐标值。

$$\begin{aligned} S1(A_i, B_j) = & \left(1 - \frac{1}{4}\sum_{k=1}^{4}\left|a_{i,k}^A - b_{j,k}^B\right|\right) * \left(1 - \left|X_{A_i} - X_{B_j}\right|\right)^{B(S_{A_i}, S_{B_j})} \times \\ & \left(\frac{2W_{A_i}W_{B_j}\left[(a_{i,1}^A + a_{i,2}^A)(b_{j,1}^B + b_{j,2}^B) + (a_{i,3}^A + a_{i,4}^A)(b_{j,3}^B + b_{j,4}^B)\right]}{W_A + W_B}\right) \times \\ & \left(\frac{1}{1 + \left[\max\left\{\left|a_{i,1}^A - b_{j,1}^B\right|, \left|a_{i,2}^A - b_{j,2}^B\right|, \left|a_{i,3}^A - b_{j,3}^B\right|, \left|a_{i,4}^A - b_{j,4}^B\right|\right\} + \left|W_{A_i} - W_{B_j}\right|\right]}\right) \end{aligned} \tag{18}$$

$$W_A=\left(W_{A_i}{}^2\right)\left[\left(a_{i,1}^A+a_{i,2}^A\right)^2+\left(a_{i,3}^A+a_{i,4}^A\right)^2\right] \tag{19}$$

$$W_B=\left(W_{B_j}{}^2\right)\left[\left(b_{j,1}^B+b_{j,2}^B\right)^2+\left(\left(b_{j,3}^B+b_{j,4}^B\right)^2\right)\right] \tag{20}$$

为了使上述相似度计算更适用于协同过滤算法的需求以提高预测准确率，本节引入信息量作为权重加入相似度计算中，梯形模糊评分信息量的定义如式(21)所示。

$$H_A^i=-r_A^i*log_2 r_A^i \tag{21}$$

对信息量做归一化处理后，将信息量作为权重改进 A_i 和 B_j 的相似度计算，具体公式如式(22)所示。

$$S(A_i,B_j)=S1(A_i,B_j)*H_A^i*H_B^j \tag{22}$$

综上所述，项目 I 和 J 的相似度如式(23)所示。

上式中 $sim_1_{i,j}$ 表示项目 i 和 j 的相似度，U 表示对项目 i 和 j 共同评分的用户集合，$n(U)$ 表示 U 的数量，若两个项目未被任何用户同时评分，则认为两个项目的相似度为 0，$R_{x,i}$ 表示用户 X 对项目 i 的梯形模糊评分，$s(R_{x,i},R_{x,j})$ 表示两个梯形模糊评分的相似度。

$$sim_1_{i,j}=\begin{cases}\dfrac{\sum_{X\in U}s\left(R_{x,i},R_{x,j}\right)}{n(U)}, & n(U)\neq 0\\ 0, & n(U)=0\end{cases} \tag{23}$$

2.3　模糊项目标签矩阵

标签作为用户和项目自身携带的数据，可以体现用户、项目的特征，提供标签维度的相似度，缓解评分稀疏性问题。除此之外，标签不受历史数据的制约，可以抑制冷启动问题。所以，引入标签改进协同过滤算法的相似度计算成为学者们研究的热点。但是，传统协同过滤算法认为标签与项目的关系只是属于和不属于，忽略了标签与项目所属关系的模糊性问题，导致预测精度较差。针对以上问题，本节利用模糊隶属度将标签属于项目的隶属度由{0,1}扩展为[0,1]，以此提升推荐质量。

推荐系统中项目包含多个标签。例如，一款手机可能包含品牌、价格、颜色和处理器型号等标签。通过这些标签反映的项目类别和特征，用户可以更高效地筛选数据。大多数推荐系统为了保证标签的专业性，会提供标签集合供项目选择，故项目与标签的关系可以使用项目标签矩阵来表示。

假设拥有 k 个标签的标签集合为 $T=\{t_1,t_2\cdots\cdots t_k\}$，其中 t_i 表示第 i 个标签，拥有 n 个项目的项目集合为 $I=\{I_1,I_2\cdots\cdots I_n\}$，其中 I_j 表示第 j 个项目，则项目标签矩阵可以被表示为 n*k 阶的矩阵 $M_{n,k}$。表 1 展示了包含 6 个项目和 5 个标签的项目标签矩阵，其中 $m_{i,j}=1$ 表示第 j 个标签属于第 i 个项目，$m_{i,j}=0$ 表示不属于。

表 1　项目标签矩阵

Tab. 1　Item label matrix

$m_{i,j}$	t_1	t_2	t_3	t_4	t_5
I_1	1	0	1	1	0
I_2	0	1	0	0	1
I_3	0	1	0	0	1
I_4	1	0	0	0	0
I_5	1	1	1	0	1
I_6	0	1	0	0	0

项目上传者和用户在分配标签时，需要根据项目内容判断标签是否属于项目，但是如何判断项目内容和标签的所属关系是一个难点。例如是否包含武打场面的电影就是功夫片，包含多少科幻元素的电影可以被定义为科幻片。所以，项目标签的隶属度存在模糊性问题，即标签与项目之间的隶属关系不应该只是属于和不属于的非此即彼关系，而应区分不同标签属于项目的程度。

项目上传者或用户对标签的使用阈值越低，则其操作的项目拥有的标签数量 Num_i 越多，因而 Num_i 与标签属于项目的隶属度 $NP_{i,j}$ 成反比。同样的，从标签角度出发，标签出现次数 $Count_i$ 与 $NP_{i,j}$ 也成反比。此外，从信息量的角度来看，利用标签计算相似度时，由于标签包含信息量的不同，故不同标签对相似度的贡献存在差异，当标签在所有项目中出现次数越多，则表示该标签包含的信息量越少，在计算基于标签的项目相似度时应该弱化该标签的影响权重。综上所述，本章算法将 Num_i 与 $Count_i$ 融合在 $NP_{i,j}$ 中。隶属度函数的定义如式(24)所示。

$$NP_{i,j}=\begin{cases}\dfrac{1}{Num_i*Count_i}, & P_{i,j}\neq 0\\ 0, & P_{i,j}=0\end{cases} \tag{24}$$

以表 1 所示的项目标签矩阵为例，通过式(24)计算项目标签隶属度 $NP_{i,j}$ 得到的模糊项目标签矩阵如表 2 所示。

表 2　模糊项目标签矩阵

Tab. 2　Fuzzy item label matrix

$NP_{i,j}$	t_1	t_2	t_3	t_4	t_5
I_1	0.111	0	0.167	0.333	0
I_2	0	0.125	0	0	0.167
I_3	0	0.125	0	0	0.167
I_4	0.333	0	0	0	0
I_5	0.083	0.063	0.125	0	0.083
I_6	0	0.25	0	0	0

为了方便后续处理，将 $NP_{i,j}$ 按照式(25)进行归一化，其中 $MAX(NP)$ 表示 NP 中最大的数值。

$$NP_{i,j}=\frac{NP_{i,j}}{MAX(NP)} \tag{25}$$

2.4　标签隶属度的相似度

构建模糊项目标签矩阵后，将每个项目的标签隶属度视作 n 维向量，通过余弦相似度计算项目相似度，计算公式如式(26)所示。

$$sim_2_{A,B}=\frac{\sum_{t\in T_A\cap T_B}NP_{A,t}*NP_{B,t}}{\sqrt{\sum_{t\in T_A}\left(NP_{A,t}\right)^2}*\sqrt{\sum_{t\in T_B}\left(NP_{B,t}\right)^2}} \tag{26}$$

2.5　项目相似度计算

将基于标签隶属度的相似度 $sim_2_{i,j}$ 和基于模糊评分的相似度 $sim_1_{i,j}$ 加权融合成项目相似度 $sim_item_{i,j}$，具体的定义如式(27)所示。λ 表示融合系数，$\lambda\in[0,1]$。

$$sim_item_{i,j}=(1-\lambda)*sim_1_{i,j}+\lambda*sim_2_{i,j} \tag{27}$$

2.6　模糊评分预测策略

传统协同过滤的评分预测策略只适用于原始评分，本小节改进该策略以适用于梯形模糊评分模型，具体步骤如下：

a) 利用模糊相似度预测用户 A 对项目 i 的模糊评分。具体方法如式(28)所示。

$$P_{A,i}=\frac{\sum_{j\in N_i}sim_item_{i,j}*A_j}{\sum_{j\in N_i}\left|sim_item_{i,j}\right|} \tag{28}$$

其中，N_i 表示项目 i 的邻域集合，A_j 表示用户 A 对项目 j 的梯形模糊评分数，$P_{A,i}$ 表示预测的梯形模糊评分数。

b) 寻找最相似的模糊评分

在经典协同过滤中，评分通常用整数表示，故需要将预测的评分四舍五入为整数。例如预测评分 4.3 将四舍五入为评分 4，在此处四舍五入的本质是为小数评分寻找最相似的整数评分。本文使用梯形模糊相似度求 $P_{A,i}$ 与用户各梯形模糊评分的相似度，为 $\tilde{P}_{A,i}$ 寻找最相似的梯形模糊评分。

c) 去模糊化

将相似度最大的梯形模糊评分 A_k 对应的原始评分 k 赋值给预测的梯形模糊评分。

$$P_{A,i}=k, if\ S_k=\max(S_1,S_2,S_3,S_4,S_5) \tag{29}$$

2.7 算法描述

本文提出了一种基于项目模糊相似度的协同过滤推荐算法，利用项目标签隶属度和梯形模糊评分确定项目的相似度，并完成推荐。

算法的具体描述如下所示。

输入：用户项目评分矩阵 R、项目标签矩阵 M、目标用户 u、目标项目 j、邻居数量 k。

输出：目标用户对目标项目的预测评分。

根据式(11)和(12)计算 $new(r_A=i)$ 和 $new(r_A)$；

a) 设置参数 t 和 p 构建梯形模糊评分模型；

b) 根据式(23)计算基于模糊评分的相似度；

c) 使用式(24)构建模糊项目标签矩阵；

d) 利用式(26)计算基于标签隶属度的相似度；

e) 设置参数 λ；

f) 根据式(27)计算项目相似度；

g) 利用式(28)预测模糊评分；

h) 使用(29)对模糊评分去模糊化。

3 实验结果及分析

本节首先介绍实验使用的数据集，然后给出评价指标，说明对比算法和实验环境，最后分析了实验结果。

3.1 实验数据集

这里采用 GroupLens 收集的 MovieLens 100K 和 1M 数据集，该数据集是验证推荐算法使用最广泛的数据集之一。数据集的相关信息如表 3 所示。

表 3　实验数据集

Tab. 3　Datasets used in experiments

数据集	用户数	项目数	评分数	稀疏度
100K	943	1682	100000	93.7%
1M	6040	3952	1000000	95.81%

3.2 评价指标

本文采用推荐系统中常用的评价指标平均绝对误差(mean absolute error，MAE)判断算法的有效性，MAE 表示预测评分与真实评分差异的平均值，MAE 越大表明预测误差越大，反之表明预测精度越高，具体定义如式(30)所示。

$$MAE=\frac{\sum_{u,i\in T}|P_{u,i}-R_{u,i}|}{|T|} \tag{30}$$

T 表示测试集，$|T|$ 表示测试集的数量，$P_{u,i}$ 表示用户 u 对项目 i 的预测评分值，$R_{u,i}$ 表示用户 u 对项目 i 的真实评分值。MAE 值越小，推荐结果的精度越高。

3.3 对比实验

为了验证本文算法 IFSCF 的有效性，对比算法有文献[13]提出的基于用户模糊相似度的协同过滤算法 FUBCF-1、文献[14]提出的改进的基于用户模糊相似度的协同过滤算法 FUBCF-2、文献[11]提出的模糊协同过滤算法 FCF、文献[17]提出的基于模糊偏差值权重的协同过滤算法 FPCF、文献[18]提出的基于模糊权重的协同过滤算法 CORFR。

3.4 实验环境

Inter(R) Core(TM) i5-9300 CPU @2.40GHz，8.0GB 内存，512GB SSD，Windows10 64 位操作系统，MatlabR2016a。

3.5 实验结果及分析

为了更清晰地表达各参数对本算法性能的影响，本节首先使用基于模糊评分的项目相似度 FRIS 作为项目相似度，分析参数 t 和 p 对 IFSCF(FRIS)算法的影响，然后调整参数 λ，将基于模糊评分的项目相似度与基于标签隶属度的项目相似度加权融合，最终比较 IFSCF 算法和对比算法在不同数据集和稀疏度中的性能差异，分析本文算法的特点。

最小满意度置信区间参数 t 决定了评分对应满意度的最小区间长度，在 5 评分系统的推荐系统中，t∈[0,0.2]。为了验证参数 t 对预测精度的影响，控制 IFSCF(FRIS)算法的其他变量固定，设步长为 0.02 进行实验，实验结果如图 3 所示。

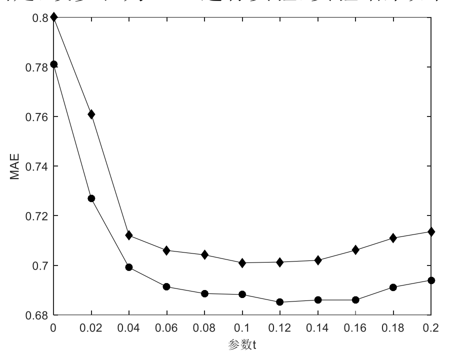


图 3　参数 t 对 MAE 值的影响

Fig. 3　Effect of parameter t on MAE

由图可知，在 100K 数据集中，随着 t 的增大，MAE 值先急速下降再缓慢回升，当 t∈(0.06,0.14)时 MAE 值稳定于 0.7 左右，当 t=0.1 时，MAE 值最小，t>0.14 时，MAE 值逐渐变大。在 1M 数据集中，MAE 值的趋势与 100K 数据集类似，MAE 值在 t=0.12 处到达最小，当 t>0.16 时出现回升现象。

MAE 值出现以上趋势的原因是评分数据过于稀疏导致评分出现概率与事实差异较大，改进前的 r_A^i 无法较精确地描述最小满意度置信区间长度，因此 t 较小时 MAE 值较大。当 0.06<t<0.16 时，改进后的 r_A^i 接近真实评分对应的满意度置信区间长度，MAE 值来到最小处。当 t 逐渐趋近于 0.2 时，评分之间失去了差异性，引起 MAE 值上升。通过以上实验和分析，固定 t 的取值为 0.1。

模糊域参数 p 表示模糊程度，决定了满意度置信区间中模糊域的占比， p∈[0,0.5]。若 p=0 则表示未使用模糊域，故模糊评分退化为原始评分，若 p=0.5 则表示满意度全是模糊域，没有确定域。为了验证 p 对实验结果的影响，在控制其他变量固定的情况下，设步长为 0.05 对 IFSCF(FRIS)算法进行实验，实验结果如图 4 所示。

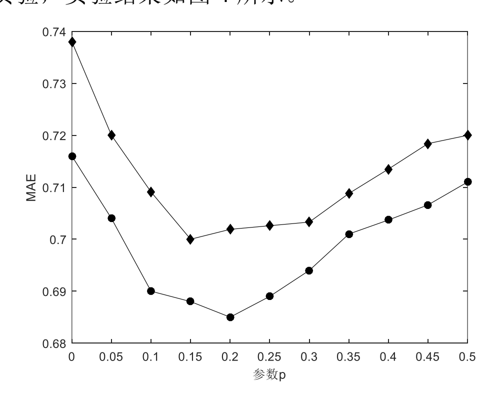


图 4　参数 p 对 MAE 值的影响

Fig. 4　Effect of parameter p on MAE

观察实验结果可知，在 100K 数据集中，随着 p 的增大，MAE 值先变小再增大，当 p=0.15 时，MAE 值最小。在 1M 数据集中，MAE 的趋势与 100K 数据集类似， MAE 值在 p=0.2 处到达最小值。

当 p 过小或过大时，评分预测的效果都不理想，这是因为当模糊域过大或者过小时，都无法较好地描述梯形模糊评分对应的满意度置信区间。由 MAE 值的最低点可知，模糊域的范围略小于确定域时效果最好。通过对比 p=0.5 和 p=0

可知，全是模糊域的满意度比全是确定域的满意度推荐质量更佳，表明了引入模糊理论构建模糊域的有效性。通过以上实验和分析，后续的实验固定 p 的取值为 0.15。

融合参数 λ 是基于模糊评分的相似度 sim_1 和基于标签的相似度 sim_2 的融合权重。为了验证参数 λ 对 IFSCF 算法性能的影响，控制近邻项目数量不变，设步长为 0.1 进行实验，实验结果如图 5 所示。

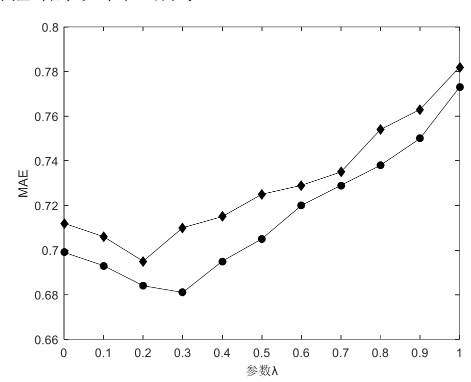


图 5　参数 λ 对 MAE 值的影响

Fig. 5　Effect of parameter　λ　on MAE

由图可知，在 100K 数据集中，当 $\lambda\in[0,0.2]$ 时，随着 λ 的增大，预测误差稳步下降并在 λ=0.2 处到达最低点，MAE 值略低于 0.7。当 $\lambda\in[0.2,1]$ 时，随着 λ 的增大，预测误差逐渐上升。数据集为 1M 时，MAE 的趋势与前者类似，当 λ=0.3 时 MAE 值低于 0.68。λ=0 和 λ=1 处的数据显示基于标签的相似度计算的评分预测精度略逊于基于模糊评分的相似度计算，这是由于前者主要负责改善项目冷启动，提高算法在稀疏数据中的效果，而后者负责提升预测精度和推荐多样性。通过以上分析，λ 的值取为 0.2。

以邻居数量为变量，取步长为 5，比较 IFSCF 算法与对比算法的预测精度，实验结果如表 4 和 5 所示，N 表示近邻数量。

表 4　100K 数据集中各算法 MAE 值的比较

Tab. 4　MAE with different neighbors(ML-100K)

N	FUBCF_1	FUBCF_2	FCF	FPCF	CORFR	IFSCF_FRIS	IFSCF
5	0.779	0.758	0.815	0.802	0.797	0.746	0.745
10	0.770	0.734	0.788	0.781	0.774	0.725	0.718
15	0.754	0.731	0.775	0.771	0.751	0.715	0.709
20	0.744	0.727	0.768	0.762	0.731	0.711	0.706
25	0.740	0.726	0.762	0.756	0.728	0.709	0.704
30	0.742	0.726	0.760	0.753	0.722	0.706	0.702
35	0.739	0.725	0.757	0.752	0.721	0.704	0.701
40	0.738	0.725	0.752	0.750	0.718	0.703	0.700
45	0.737	0.726	0.752	0.749	0.717	0.703	0.700
50	0.737	0.729	0.75	0.749	0.715	0.702	0.699

表 5　1M 数据集中各算法 MAE 值的比较

Tab. 5　MAE with different neighbors(ML-1M)

N	FUBCF_1	FUBCF_2	FCF	FPCF	CORFR	IFSCF_FRIS	IFSCF
5	0.765	0.740	0.801	0.783	0.803	0.731	0.745
10	0.758	0.721	0.772	0.763	0.775	0.706	0.709
15	0.739	0.714	0.760	0.755	0.741	0.699	0.701
20	0.733	0.711	0.748	0.751	0.725	0.695	0.693
25	0.730	0.709	0.745	0.745	0.721	0.691	0.690
30	0.728	0.705	0.745	0.740	0.719	0.689	0.686
35	0.725	0.705	0.741	0.736	0.716	0.687	0.683
40	0.721	0.703	0.742	0.733	0.712	0.686	0.682
45	0.720	0.702	0.741	0.732	0.711	0.685	0.682
50	0.719	0.702	0.740	0.731	0.711	0.685	0.681

由表可知，在 100K 数据集中，任意邻域数量下 IFSCF 算法的精度皆高于对比算法，在 1M 数据集中，当邻域数量较小时，IFSCF 算法的误差大于 FUBCF-2 和 IFSCF(FRIS)算法，但当邻域数量扩大后，IFSCF 算法的误差逐渐变为最小。

IFSCF 算法引入了标签模糊隶属度，理论上可以更好地应对评分稀疏性问题。为了验证以上推测，本实验使用 100K 数据集，在保证用户数和项目数不变的前提下，逐步减少数据集中的评分数量，将数据集的稀疏度从 0.937 逐渐提升到 0.99，比较 IFSCF 算法与对比算法在不同稀疏度中的表现。实验结果如图 6 所示。

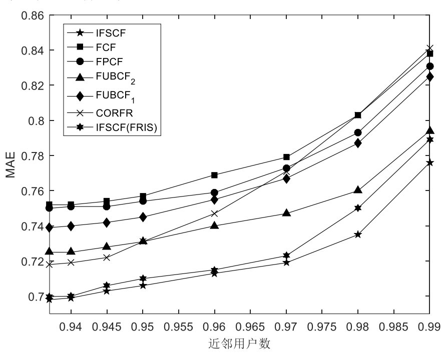


图 6　不同稀疏度下各算法 MAE 的比较

Fig. 6　MAE with different sparsity

从实验结果可知，随着稀疏度的增大，可使用的数据逐渐减少，各算法的 MAE 值都不同程度地增大，并且在稀疏度大于 97%后增速变快。其中，IFSCF 算法的 MAE 值增幅较小，预测精度最高，可在稀疏数据中较好地完成推荐。

4　结束语

本文引入模糊理论改善了评分-满意度和项目-标签的模糊性问题，提出了一种基于项目模糊相似度的协同过滤推荐算法，利用梯形模糊数描述评分与满意度的映射关系，改进模糊相似度计算策略以提升相似度计算的精度，使用隶属度函数判断标签与项目的所属程度，根据项目标签隶属度向量计算基于标签的相似度，改进评分预测策略以进行模糊评分估计，对基于项目的协同过滤算法全过程实施了模糊处理。实验结果表明，该算法可在一定程度上缓解模糊性问题并可改善评分数据稀疏性带来的不利影响。

参考文献：

[1] 魏甜甜，陈莉，范婷婷，等. 结合项目流行度加权的协同过滤推荐算法 [J]. 计算机应用研究, 2020, 37 (3): 676-679. (Wei Tiantian, Chen Li, Fan Tingting, *et al.* Collaborative filtering recommendation algorithm based on item popularity weighting [J]. Application Research of Computers, 2020, 37 (3): 676-679.)

[2] Ekstrand M D. Collaborative filtering recommender systems [J]. Acm Transactions on Information Systems, 2007, 22 (1): 5-53.

[3] Wei Jian, He Jianhua, Chen Kai, *et al.* Collaborative filtering and deep learning based recommendation system for cold start items [J]. Expert Systems With Applications, 2017, 69 (69): 29-39.

[4] Hu Yan, Shi Weisong, Li Hong, *et al.* Mitigating data sparsity using similarity reinforcement-enhanced collaborative filtering [J]. Acm Transactions on Internet Technology, 2017, 17 (3): 1-20.

[5] Yera R, Martinez L. Fuzzy tools in recommender systems: a survey [J]. International Journal of Computational Intelligence Systems, 2017, 10 (1): 776-803.

[6] Kant S, Mahara T, Jain V K, *et al.* Fuzzy logic based similarity measure for multimedia contents recommendation [J]. Multimedia Tools and Applications, 2019, 78 (4): 4107-4130.

[7] Berkani L. Social-Based collaborative recommendation: bees swarm optimization based clustering approach [C]// International Conference on Model and Data Engineering. Springer, Cham, 2019: 156-171.

[8] Tsai C H. A fuzzy-based personalized recommender system for local businesses [C]// Proceedings of the 27th ACM Conference on Hypertext and Social Media. ACM, 2016: 297-302.

[9] Vashisth P, Khurana P, Bedi P. A fuzzy hybrid recommender system [J]. Journal of Intelligent & Fuzzy Systems, 2017, 32 (06): 3945-3960.

[10] Wasid M, Kant V. A particle swarm approach to collaborative filtering based recommender systems through fuzzy features [J]. Procedia Computer Science, 2015, 54 (06): 440-448.

[11] Kant V, Bharadwaj K K. Enhancing recommendation quality of content-based filtering through collaborative predictions and fuzzy similarity measures [J]. Procedia engineering, 2012, 38 (3): 939-944.

[12] Zhang Xixiang, Ma Weimin, Chen Liping. New similarity of triangular fuzzy number and its application [J]. The Scientific World Journal, 2014, 2014 (1): 1-7.

[13] 吴毅涛，张兴明，王兴茂，等. 基于用户模糊相似度的协同过滤算法 [J]. 通信学报，2016, 37 (1): 198-206. (Wu Yitao, Zhang Xingming, Wang Xingmao, *et al.* User fuzzy similarity-based collaborative filtering recommendation algorithm [J]. Journal on Communications. 2016, 37 (1): 198-206.)

[14] Wu Yitao, Zhang Xingming, Hong Yyu, *et al.* Collaborative filtering recommendation algorithm based on user fuzzy similarity [J]. Intelligent Data Analysis, 2017, 21 (2): 311-327.

[15] L. A. Zadeh. Fuzzy Sets [J]. Information & Control, 1965, 8 (3): 338-353.

[16] Ahmad S A S, Mohamad D, Sulaiman N H, *et al.* A distance and set theoretic-based similarity measure for generalized trapezoidal fuzzy numbers [C]// AIP Conference Proceedings. AIP Publishing LLC, 2018, 1974 (1): 02-43.

[17] Al-Shamri M Y H, Al-Ashwal N H. Fuzzy-weighted similarity measures for memory-based collaborative recommender systems [J]. Journal of Intelligent Learning Systems and Applications, 2014, 6 (01): 1-10.

[18] Lee S. Collaborative filtering using fuzzy rank-based similarity measures [C]// 2018 International Conference on Control, Artificial Intelligence, Robotics & Optimization (ICCAIRO) . IEEE, 2018: 84-89.